UNITED STATES PATENT APPLICATION FOR:

METHOD AND APPARATUS FOR RECOGNIZING SPEECH IN A NOISYENVIRONMENT

INVENTOR:

VENKATA RAMANA RAO GADDE

ATTORNEY DOCKET NUMBER: SRI/4544-2

CERTIFICATION OF MAILING UNDER 37 C.F.R. 1.10

I hereby certify that this New Application and the documents referred to as enclosed therein are being deposited with the United States Postal Service on August 15, 2001, in an envelope marked as "Express Mail United States Postal Service", Mailing Label No. EL849340783US, addressed to: Box Patent Application, Assistant Commissioner for Patents, Washington, D.C. 20231.

Signature

KIN-WAH TONG

Name

8/15/01

Date of signature

THOMASON, MOSER & PATTERSON LLP
595 Shrewsbury Ave.
Shrewsbury, New Jersey 07702
(732)530-9404

METHOD AND APPARATUS FOR RECOGNIZING SPEECH IN A NOISY ENVIRONMENT

The present invention relates to an apparatus and concomitant method for audio signal processing. More specifically, the present invention provides a new noise compensation method for adapting speech models to noise in a recognition system, thereby improving the speed of speech recognition and reducing computational cycles.

BACKGROUND OF THE DISCLOSURE

Speech recognition systems are designed to undertake the difficult task of extracting recognized speech from an audio signal, e.g., a natural language signal. The speech recognizer within such speech recognition systems must account for diverse acoustic characteristics of speech such as vocal tract size, age, gender, dialect, and the like. Artificial recognition systems are typically implemented using powerful processors with large memory capacity to handle the various complex algorithms that must be executed to extract the recognized speech.

To further complicate the complex speech recognition process, the audio signal is often obtained or extracted from a noisy environment, e.g., an audio signal captured in a moving vehicle or in a crowded restaurant, thereby compromising the quality of the input audio signal. To address the noisy background or environmental contamination, the speech recognizer can be implemented with various noise compensation algorithms.

Noise compensation schemes include the Parallel Model Combination (PMC) and other model adaptation techniques. However, these schemes often require large amounts of memory and are computationally intensive. To illustrate, the PMC method is a method of adding and synthesizing a Hidden Markov Model (HMM) (speech HMM) learned by speech collected and recorded in a noiseless environment and an HMM (noise HMM) learned by noise. In the noise process of the PMC, it is presumed that additiveness of noise and speech is established in a linear spectrum region. In contrast, in the HMM, parameters of a logarithm spectrum system, such as a cepstrum and the like, are often used as a characteristic amount of the speech. According to the PMC method, those parameters are converted into

the linear spectrum region and then are added and synthesized in the linear spectrum region of the characteristic amount, which is derived from the speech HMM and noise HMM. After the speech and the noise are synthesized, an inverse operation is performed to return the synthesized value from the linear spectrum region to the cepstrum region, thereby obtaining a noise superimposed speech HMM. However, although the PMC is effective in addressing additive noise, the PMC method is very computationally expensive because the nonlinear conversion is executed to all of the models. Namely, the amount of calculations is very large, the processing time is very long, and it may not be suitable for a real time application or a portable application where processing resources and memory capacity are limited.

Therefore, a need exists for a fast and computationally inexpensive method that addresses the problem of speech recognition in noisy environments without the need of any prior recognition pass or large memory capacity.

SUMMARY OF THE INVENTION

The present invention is an apparatus and a concomitant method for speech recognition. In one embodiment, the present method is referred to as a "Dynamic Noise Compensation" (DNC) method where the novel method estimates the models for noisy speech using models for clean speech and a noise model. Specifically, the model for the noisy speech is estimated by interpolation between the clean speech model and the noise model. In practice, the noise model is approximated by a noise estimate from the noisy speech. This novel approach reduces computational cycles and does not require large memory capacity. These significant savings allow the present invention to be implemented in a real time application and/or a portable application, e.g., where the speech recognition system is a portable device.

BRIEF DESCRIPTION OF THE DRAWINGS

The teachings of the present invention can be readily understood by considering the following detailed description in conjunction with the accompanying drawings, in which:

FIG. 1 illustrates a block diagram of a speech recognition system of the present invention;

FIG. 2 illustrates a block diagram of a generic speech recognizer;

FIG. 3 illustrates a block diagram of a speech recognizer of the present invention;

FIG. 4 illustrates a block diagram of a dynamic noise compensation module of the present invention; and

FIG. 5 illustrates a block diagram of a speech recognition system of the present invention as implemented using a general purpose computer.

To facilitate understanding, identical reference numerals have been used, where possible, to designate identical elements that are common to the figures.

DETAILED DESCRIPTION

FIG. 1 illustrates a block diagram of a speech recognition device or system 100 of the present invention. In one embodiment, the speech recognition device or system 100 is implemented using a general purpose computer or any other hardware equivalents as shown in FIG. 5 below. Although the recognition device or system 100 is preferably implemented as a portable device, it should be noted that the present invention can also be implemented using a larger computer system, e.g., a desktop computer or server and the like .

The speech recognition device or system 100 comprises a sampling and Analog-to-Digital (A/D) conversion module 110, a feature extractor or feature extraction module 120, a speech recognizer or a speech recognizer module 130 and various Input/Output (I/O) devices 140. In operation, an input audio signal (e.g., a speech signal) on path 102 is received by the sampling and Analog-to-Digital (A/D) conversion module 110, where the input signal is sampled and digitized from a microphone (not shown) into a sequence of samples that are later processed by a processor.

The digitized sequence of samples is then forwarded on path 103 to the feature extraction module 120. The sample sequence is first grouped into frames (commonly 1 centi-second in length) and speech features are extracted for each of the frames using various signal processing methods. Some examples of these are Mel-cepstral features, or PLP cepstral features.

Specifically, conventional feature extraction methods for automatic speech recognition generally rely on power spectrum approaches, whereby the acoustic signals are generally regarded as a one dimensional signal with the assumption that the frequency content of the signal captures the relevant feature information. This is the case for the spectrum representation, with its Mel or Bark variations, the cepstrum, FFT-derived (Fast Fourier Transform) or LPC-derived (Linear Predictive Coding), LPC derived features, the autocorrelation, the energy content, and all the associated delta and delta-delta coefficients.

Cepstral parameters are effectively used for efficient speech and speaker recognition. Originally introduced to separate the pitch contribution from the rest of the vocal cord and vocal tract spectrum, the cepstrum has the additional advantage of approximating the Karhunen-Loeve transform of speech signal. This property is highly desirable for recognition and classification. In one embodiment of the present invention, the speech features on path 104 can be Mel-cepstral features, or PLP cepstral features.

It should be noted that the present invention is not limited to a particular type of feature, as long as the same features are used to train the models and used during the recognition process. Namely, the present invention is not feature dependent.

In turn, the speech recognizer 130 receives the speech features and is able to decode the "recognized text" from the speech features using various models as discussed below. Finally, the recognized text on path 105 is further processed by various I/O devices or other processing modules 140, e.g., natural language processing module, speech synthesizer and the like.

FIG. 2 illustrates a block diagram of a generic speech recognizer 130 comprising a text decoder or extractor 210, acoustic models 220 and a language model 230. Specifically, the input speech features on path 104 obtained from the utterance (input audio signal) are decoded using the acoustic models 220 and a language model 230. The acoustic models are trained using a large amount of training speech. Typically, acoustic models are Hidden Markov Models (HMMs) trained for each sound unit (phone, triphone, etc.). Each HMM usually has 3 states and each state may be modeled using one or more gaussians. Some of the states may be tied by

sharing the same gaussians. The HMM techniques are used to identify the most likely sequence of words that could have produced the speech signal.

However, one problem with the HMM based speech recognition is the mismatch between the speech data used for training and during testing/use. Typical training data is obtained under controlled environments that are noise free. However, the test speech is obtained in real world conditions which are usually noisy. This mismatch leads to a significant loss in performance. Thus, the present DNC is developed to compensate for the mismatch.

FIG. 3 illustrates a block diagram of a speech recognizer 130 of the present invention comprising a text decoder or extractor 210, a dynamic noise compensator, or a dynamic noise compensation module 310, clean acoustic models 320 and a language model 230. FIG. 3 illustrates the speech recognizer using the DNC of the present invention. In one embodiment, the input noisy speech features are used to compensate the clean speech models (using the DNC formula as disclosed below) to generate models for noisy speech. These models are then used along with the language model 230 to decode the input speech features on path 104.

FIG. 4 illustrates a block diagram of the Dynamic Noise Compensation module 310 of the present invention. It should be noted that FIG. 4 when viewed with the discussion provided below, also serves as a flowchart for the present noise compensation method.

FIG. 4 illustrates the architecture of the DNC comprising a noise estimation module 410, a model weight selection module 420, two multipliers 430 and a summer 440. The first two stages are the noise model estimation module and the model weight selection module. Specifically, the noise model is estimated using the features corresponding to the noise in the input. In one implementation, the energy is used to identify the low energy frames. The noise estimate is then used to select appropriate weight for the interpolation. This weight is then used to combine the clean speech models and the noise model to generate the models for noisy speech.

Specifically, the noise energy estimate is used to compute an estimate of the signal to noise ratio (SNR). In one implementation, the SNR is approximated by the ratio of the maximum energy to the estimated noise

energy. This SNR is used to look up a table of SNR-Weight pairs and the weight corresponding to the closest SNR value in the table is used.

In one embodiment, the SNR-Weight table is generated in accordance with the following procedure. First, the clean speech is used to build the clean speech HMMs. Second, a test set of clean speech is used and corrupted using random samples of a variety of noises (for example, car noise or other noises in an environment that the speech recognition system is intended to operate within). The noise energy is then changed to produce noisy speech data at different SNRs. The present DNC algorithm is then applied with a number of weights, where the appropriate weight is then selected (i.e., the weight which produced the best recognition performance for a noisy speech having a particular SNR). This estimation is repeatedly performed at different SNRs, thereby generating the table of SNR-Weight pairs.

Namely, the Dynamic Noise Compensation is a new method that estimates the models for noisy speech using models for clean speech and a noise model. Current state-of-the-art speech recognition systems use HMMs to model speech units like triphones. A typical HMM has 3 states each modeling the initial, middle and the final segments of that triphone. Typically, these models are Gaussian Mixture Models (GMMs) which are a collection of gaussians modeling the probability distribution of the features belonging to that state. Each gaussian is represented by two parameters, the mean and the variance. The use of HMMs in the field of speech recognition is well known and description of HMMs can be found in general references such as L. Rabiner and B. Juang, "Fundamentals of speech recognition", Prentice Hall, 1993 and Frederick Jelinek, "Statistical Methods for Speech Recognition", MIT press, Cambridge, MA, 1998.

In the context of the present DNC, the HMMs are trained using clean speech data. The training procedure estimates the parameters of all the gaussians in the models. In DNC, these parameters are modified so that they now model noisy speech.

Consider a gaussian modeling clean speech. Let the mean of the gaussian be M and standard deviation C. If the noise estimate from the noisy speech is N, then the mean M' and variance C' for noisy speech are estimated as:

$$M' = W*M + (1-W)*N,\ 0 < W < 1 \qquad (1)$$
$$C' = C$$

The interpolation weight W is determined from an estimate of the Signal to Noise Ratio (SNR). In one embodiment, the noise estimate (and the SNR) is obtained by averaging low energy frames in the input noisy speech. Specifically, to estimate the noise, the frames with the lowest energy in the input speech are identified. These frames are assumed to be noise frames and these are used to estimate a noise model. Generally, the noise model can be a GMM (i.e., a mixture of gaussians), but in practice it has been found that a single gaussian model of noise works quite well. In turn, the mean of the noise model (N) is used in the DNC formula to estimate the noisy speech models. This noise estimate is used to update all the gaussians in the clean speech models (HMMs) using the above formula.

FIG. 5 illustrates a block diagram of a speech recognition system 500 of the present invention as implemented using a general purpose computer. The speech recognition device or system 500 comprises a processor (CPU) 512, a memory 514, e.g., random access memory (RAM) and/or read only memory (ROM), a speech recognizer module 516, and various input/output devices 520, (e.g., storage devices, including but not limited to, a tape drive, a floppy drive, a hard disk drive or a compact disk drive, a receiver, a transmitter, a speaker, a display, a speech signal input device, e.g., a microphone, a keyboard, a keypad, a mouse, an A/D converter, and the like).

Namely, speech recognizer module 516 can be the speech recognizer module 130 of FIG. 1. It should be understood that the speech recognizer module 516 can be implemented as a physical device that is coupled to the CPU 512 through a communication channel. Alternatively, the speech recognizer module 516 can be represented by one or more software applications (or even a combination of software and hardware, e.g., using application specific integrated circuits (ASIC)), where the software is loaded from a storage medium, (e.g., a magnetic or optical drive or diskette) and operated by the CPU in the memory 514 of the computer. As such, the speech recognizer module 516 (including associated methods and data structures) of the present invention can be stored on a computer readable

medium, e.g., RAM memory, magnetic or optical drive or diskette and the like. Additionally, it should be understood that various modules and models (e.g., feature extraction module, language models, acoustic models, speech synthesis module, translation module and its sub-modules) as discussed above or known in the art can be stored and recalled into memory 514 for execution.

Although various embodiments which incorporate the teachings of the present invention have been shown and described in detail herein, those skilled in the art can readily devise many other varied embodiments that still incorporate these teachings.